Nucleotide Sequence Manipulation
IMBB Workshop
14, May 2013

Joyce Njoki Nzioki
BecA-ILRI Hub, Nairobi, Kenya
http://hub.africabiosciences.org/
http://www.ilri.org/
j.n.juguna@cgiar.org
Any analytical methods connected to identifying higher biological meaning out of raw sequence data.

Annotation: bridging the gap from sequence to the biology of the organism.
Nucleotide Level Analysis

- Sequence transcription (Dna to Rna)
- Reverse complement
- Finding regions of interest
- Translation DNA in 6 reading frames
- Database searching
Converting DNA sequences to RNA

- DNA is a double stranded, one strand is designated “forward” the other “reverse” or (+ and - strand).

- Typically a sequence is read from direction 5’-3’ for the forward strand that means left to right and the reverse strand right to left.

- During transcription, RNA polymerase moves along the DNA in a 5’-3’ orientation producing a complementary strand which includes **uracil** in place of **thymidine**.
Nucleotide sequence Analysis

• In the DNA double helix Adenine pairs with thymine and guanine with cytosine.
  - A and T connected with two hydrogen bonds.
  - C and G connected with three hydrogen bonds.

• Each base pair in the double helix structure must pair with its complementary strand it is possible to determine the sequence of the complementary strand.

  GATTCGTACG
  CTAAGCATGC

There are tools available for complementary strand determination: CLC main workbench, Expasy tools.
Reverse Compliment

In some cases you may need the reverse compliment of a DNA sequence: (this is the complementary strand in the 5’- 3’ direction)

1. For example as a negative control when doing a DNA hybridization experiment

There are a number of available tools for this:

• CLC main workbench (to be done in practical session).
• EMBOSS reveseq (reverse and complement a sequence)
  http://emboss.bioinformatics.nl/cgi-bin/emboss/revseq
Complementary strand

- On CLC to get a complementary strand of a given sequence:
  - Go to the Tool box menu | Nucleotide analysis | select the Reverse Complement Sequence option.
  - To reverse a sequence from the 5’ -3’ or 3’ – 5’ go to the Toolbox menu | Nucleotide analysis | select Reverse Sequence option.
Standard Genetic Code

• Genes are stretches of DNA that encode information for building proteins.
• A triplet of nucleotides codes for an amino acid the building blocks of proteins.
Translating nucleotide sequences

- DNA codes for amino acids a three letter genetic code.
- Translation of DNA to Amino acids can be done in 6 possible reading frames.

**GATTCGTACG**
**CTAAGCATGC**

1. GAT TCG TAC
   - Asp Ser Thr

2. GATT CGT ACG
   - Ile Arg Thr

3. GATTC GTACG
   - Phe Val

4. CTA AGC ATG C
   - Leu Ser Met

5. C TAA GCA TGC
   - # Ala Arg

6. CT AAG CAT GC
   - Lys His
Translation tools

There are various computational tools available for sequence translation:

1. CLC main workbench: (Practicals in the next session).
3. EMBOSS: transeq (translate nucleic acids)
4. SeWeR analysis: (Sequence analysis using web resources) http://www.bioinformatics.org/SeWeR/
Gene finding refers to identifying stretches of sequences (genes) in genomic DNA that are biologically functional.
Prokaryotes:
1. Small genomes ~ 5 million bp in bacteria, in this case gene finding is a matter of identifying long ORFs.
2. E.g. Haemophilus influenzae ~85% of its genome is in coding regions
3. Calling genes involves using programs that carry out 6 frame translation and identify all the ORFs longer than a given threshold.

Eukaryotes:
1. Larger genomes ~3 Billion bases
2. for yeast, <25% of its genome is in coding regions, for humans it is less than <3%.
3. Splicing and alternative splicing
4. Introns and exons present
Gene:

Start of transcription

Gene:

Exon 1  Intron 1  Exon 2  Intron 2  Exon 3

Transcription

Primary transcript:

Exon 1  Intron 1  Exon 2  Intron 2  Exon 3

Splicing

Mature transcript:

Exon 1  Exon 2  Exon 3
Gene finding Approaches

Computations methods proposed to find genes in both prokaryotes and eukaryotes include.

- Homology based approaches
- *Ab Initio* approaches
- Comparative genomics
Homology based Approaches

- Based on **sequence similarity** eg BLAST
- Searching a library of other organisms we identify known genes that resemble the target sequence.
- If the identified sequences are genes, the target sequence is putatively a gene and function can be conferred.
- **Limitation:** identifies only genes that are protein coding and are present in databases
Ab Initio Approaches

• Rely on statistical qualities of exons rather than homology.
• **Tools:** GeneScan, fgenish, HMMER, Augustus
• **Limitation:**
  - Prone to many false positives
  - Often can’t predict splice variants
  - Cannot predict 3’ and 5’ UTRs
Compilation of a catalogue of proteins of the organism, name them and assign putative function.

Protein family classification uses similarities to better characterized proteins. Search for similarity using Blastp or PSI-Blast tools against several protein databases
Protein level annotation

- **Domain searches**: Protein domains are functional units of a protein; these are key in predicting function of genes.
- There are several databases of functional domains that can be searched.
- PFAM, PRINTS, PROSITE, ProDom, BLOCKS, SMART.
- InterPro integrates all the protein signatures and domains in one resource.
Process level annotation

How do the of genes and proteins relate to.
  Cell cycle
  Apoptosis
  Metabolism
  Health and disease
  Resistance
  Reproduction
Gene Ontology (GO)

A standard vocabulary for describing function of Eukaryotic genes. Provides vital clues to the role that genes and proteins have in biological process and provide a rich layer to annotation.

- **Molecular function**: - task carried out by individual genes e.g enzymatic activity
- **Biological process**: used for broader biological roles e.g meiosis.
- **Cellular component**: describes genes in terms of subcellular structure e.g localized to nucleus
Conclusion

Process level annotation extends beyond purely computational work.

Bench research, nucleotide level, protein level and process level annotations allow scientists to obtain functional information for novel gene products.
THE END
PRACTICAL STARTS SHORTLY